



TITLE:

Patterns and ecological drivers of ocean viral communities.

AUTHOR(S):

Brum, Jennifer R; Ignacio-Espinoza, J Cesar; Roux, Simon; Doulcier, Guilhem; Acinas, Silvia G; Alberti, Adriana; Chaffron, Samuel; ... Wincker, Patrick; Karsenti, Eric; Sullivan, Matthew B

CITATION:

Brum, Jennifer R ...[et al]. Patterns and ecological drivers of ocean viral communities.. Science 2015, 348(6237): 1261498.

ISSUE DATE:

2015-05-22

URL:

<http://hdl.handle.net/2433/197954>

RIGHT:

This is the author's version of the work. It is posted here by permission of the AAAS for personal use, not for redistribution. The definitive version was published in [Patterns and ecological drivers of ocean viral communities] on Vol.348 no.6237 DOI:10.1126/science.1261498; この論文は出版社版ではありません。引用の際には出版社版をご確認ください。 ; This is not the published version. Please cite only the published version.

Title: Patterns and ecological drivers of ocean viral communities

Authors: Jennifer R. Brum^{§,1}, J. Cesar Ignacio-Espinoza^{§,2}, Simon Roux^{§,1}, Guilhem Doulier^{1,3}, Silvia G. Acinas⁴, Adriana Alberti⁵, Samuel Chaffron^{6,7,8}, Corinne Cruaud⁵, Colombar de Vargas^{9,10}, Josep M. Gasol⁴, Gabriel Gorsky^{11,12}, Ann C. Gregory¹³, Lionel Guidi^{11,12}, Pascal Hingamp¹⁴, Daniele Iudicone¹⁵, Fabrice Not^{9,10}, Hiroyuki Ogata¹⁶, Stephane Pesant^{17,18}, Bonnie T. Poulos¹, Sarah M. Schwenck¹, Sabrina Speich^{19,†}, Celine Dimier^{9,10,20}, Stefanie Kandels-Lewis^{21,22}, Marc Picheral^{11,12}, Sarah Searson^{11,12}, Tara Oceans Coordinators[‡], Peer Bork^{21,23}, Chris Bowler²⁰, Shinichi Sunagawa²¹, Patrick Wincker^{5,24,25}, Eric Karsenti^{20,22,*}, & Matthew B. Sullivan^{1,2,13,*}

[§]co-first authors

Affiliations:

¹ Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona, 85721, USA

² Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona, 85721, USA

³ Environmental and Evolutionary Genomics Section, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS, UMR8197, INSERM U1024, 75230 Paris, France

⁴ Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)-CSIC, Pg. Marítim de la Barceloneta 37-49, Barcelona, E08003, Spain

⁵ CEA - Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057, Evry, France

⁶ Department of Microbiology and Immunology, Rega Institute KU Leuven, Herestraat 49, 3000 Leuven, Belgium

⁷ Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium

⁸ Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

⁹ CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

¹⁰ Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

¹¹ CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France

¹² Sorbonne Universités, UPMC Univ Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-mer, France

¹³ Soil, Water, and Environmental Science, University of Arizona, Tucson, Arizona, 85721, USA

- 37 ¹⁴ Aix Marseille Université CNRS IGS UMR 7256 13288, Marseille, France
- 38 ¹⁵ Stazione Zoologica Anton Dohrn, Villa Comunale, 80121, Naples, Italy
- 39 ¹⁶ Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-001, Japan
- 40 ¹⁷ PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen,
41 Bremen, Germany
- 42 ¹⁸ MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen,
43 Germany
- 44 ¹⁹ Laboratoire de Physique des Océan UBO-IUEM Palce Copernic 29820 Polouzané, France
- 45 ²⁰ Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and
46 CNRS UMR 8197, Paris, F-75005, France
- 47 ²¹ Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr.
48 1, 69117 Heidelberg, Germany
- 49 ²² Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1 69117
50 Heidelberg, Germany
- 51 ²³ Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany
- 52 ²⁴ CNRS, UMR 8030, CP5706, Evry, France
- 53 ²⁵ Université d'Evry, UMR 8030, CP5706, Evry, France
- 54 † Current address: Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD),
55 Ecole Normale Supérieure, 24 rue Lhomond 75231 Paris, Cedex 05, France
- 56 ‡ *Tara* Oceans coordinators and affiliations are listed at following the Acknowledgments.
- 57 *Correspondence to: mbsulli@gmail.com, karsenti@embl.de

Abstract: Viruses influence ecosystems by modulating microbial population size, diversity, metabolic outputs, and gene flow. Here we use quantitative double-stranded DNA (dsDNA) viral-fraction metagenomes (viromes) and whole viral community morphological datasets from 43 *Tara* Oceans expedition samples to assess viral community patterns and structure in the upper ocean. Protein cluster cataloging defined pelagic upper-ocean viral community pan and core gene sets and suggested this sequence space is well-sampled. Analyses of viral protein clusters, populations, and morphology revealed biogeographic patterns whereby viral communities were passively transported on oceanic currents and locally structured by environmental conditions that impact host community structure. Together these investigations establish a global ocean dsDNA viromic dataset with analyses supporting the seed-bank hypothesis to explain how oceanic viral communities maintain high local diversity.

One Sentence Summary: Global patterns that emerge from the *Tara* Oceans Virome dataset support a seed-bank structure underlying observed biogeography in ocean viral communities.

Main Text: Ocean microbes produce half of the oxygen we breathe (1) and drive much of the substrate and redox transformations that fuel Earth's ecosystems (2). However, they do so in a constantly evolving network of chemical, physical and biotic constraints – interactions which are only beginning to be explored. Marine viruses are presumably key players in these interactions (3, 4) as they affect microbial populations through lysis, reprogramming of host metabolism, and horizontal gene transfer. Here we strive to develop an overview of ocean viral community patterns and ecological drivers.

The *Tara* Oceans expedition provided a platform for sampling ocean biota from viruses to fish larvae with comprehensive environmental context (5). Prior virus-focused work from this expedition has helped optimize the dsDNA viromic sample-to-sequence workflow (6), evaluate ecological drivers of viral community structure as inferred from morphology (7), and map ecological patterns in the large dsDNA nucleo-cytoplasmic viruses using marker genes (8). Here we explore global patterns and structure of ocean viral communities using 43 samples from 26 stations in the *Tara* Oceans expedition (Supplementary File S1) to establish dsDNA viromes from viral-fraction (<0.22 µm) concentrates and quantitative whole viral community morphological datasets from unfiltered seawater. Viruses lack shared genes that can be used for investigation of community patterns. Therefore, we used three levels of information to study such patterns: (i) protein clusters (PCs, 9) as a means to organize virome sequence space commonly dominated by unknown sequences (63–93%, 10), (ii) populations, using established metrics for viral contig recruitment (11), and (iii) morphology, using quantitative transmission electron microscopy (qTEM, 7).

The *Tara* Oceans Viromes (TOV) dataset

The 43 *Tara* Oceans Viromes (TOV) dataset is comprised of 2.16 billion ~101-bp paired-end Illumina reads (Supplementary File 1), largely representing epipelagic ocean viral communities (only 1 of 43 viromes are from mesopelagic waters, Environment Ontology feature ENVO:00000213) from the surface (ENVO:00002042) and deep chlorophyll maximum (DCM;

ENVO:01000326) throughout seven oceans and seas (Supplementary File S1). The TOV dataset offers deeper sampling of surface ocean viral communities, but under-represents the deep ocean relative to the Pacific Ocean Viromes dataset (POV, 10) which includes 16 viromes from aphotic zone waters. In all viromes, sampling and processing affects what viruses are represented (6, 12-14). We filtered TOV seawater samples through 0.22 μm pore-sized filters and then concentrated viruses in the filtrate using iron chloride flocculation (15). These steps would have removed most cells, but also excluded any viruses larger than 0.22 μm . We then purified the resulting TOV viral concentrates using DNase treatment, which is as effective as density gradients for purifying ocean viral concentrates (14). This DNase-only step is unlikely to impact viral representation in the viromes, but reduces non-viral DNA contamination. Finally, we extracted DNA from the samples and prepared sequence libraries using linker amplification (13). These steps preserve quantitative representation of dsDNA viruses in the resulting viromes (12, 13), but the ligation step excludes RNA viruses, and is biased against single-stranded DNA (ssDNA) viruses (12).

We additionally applied qTEM (7) to paired whole seawater samples to evaluate patterns in whole viral communities. This method simultaneously considers ssDNA, dsDNA, and RNA viruses, though without knowledge of their relative abundances since particle morphology does not identify nucleic acid type. In the oceans, total virus abundance estimates based on TEM analyses, which include all viral particles, are similar to estimates based on fluorescent staining, which inefficiently stains ssDNA and RNA viruses (16-24). This suggests that most ocean viruses are dsDNA viruses. However, one study quantifying nucleic acids at a single marine location suggests RNA viruses may constitute as much as half of the viral community there (16). It remains unknown what the relative contribution of these viral types is to the whole viral community, but our analyses suggest small dsDNA viruses likely dominate as follows. The viromes capture the <0.22 μm dsDNA viruses of bacteria and archaea that are thought to dominate marine viral communities, whereas qTEM analysis includes all viruses regardless of size, nucleic acid type, or host (7). In these whole seawater samples used for qTEM, we found that viral capsid diameters ranged from 26 to 129 nm, with the per-sample average capsid diameter constrained at 46–66 nm (Fig. 1). We detected no viral particles larger than 0.22 μm among 100 randomly counted particles from each of 41 qTEM samples. These findings are similar to those from a subset of these *Tara* Oceans stations (14 of the 26 stations; 7), and indicate that size fractionation using 0.22 μm filtration to prepare viromes did not substantially bias the TOV dataset.

TOV Protein Clusters for Comparison of Local and Global Genetic Richness and Diversity

Across the 43 viromes, a total of 1,075,763 PCs were observed, with samples beyond the 20th virome adding few PCs (Fig. 2A). When combining TOV with 16 photic-zone viromes from the POV dataset (10), the number of PCs increased to 1,323,921, but again approached a plateau (Fig. 2B). These results suggest that, while impossible to sample completely, the sequence space corresponding to dsDNA viruses from the epipelagic ocean is now relatively well sampled. This contrasts results from marine microbial metagenomic surveys using older sequencing

technologies (9), but is consistent with those from this expedition (25), as well as findings from viral sequence datasets which suggest a limited range of functional diversity derived from bacterial and archaeal viral isolates (26) and the POV dataset (27).

PCs were next used to establish the core genes shared across the TOV dataset (Fig. 2A). Broadly, there were 220, 710 and 424 core PCs shared across all surface and DCM viromes, surface viromes only, and DCM viromes only, respectively. The number of core PCs in the upper-ocean TOV samples (220 PCs) was thus less than the number of photic-zone core PCs in POV (565 PCs; 28), likely because the POV dataset includes only the Pacific Ocean while TOV includes samples from seven oceans and seas. However, the number of core PCs in the upper-ocean TOV samples exceeded the total number of core PCs observed in POV (180 PCs; 28), likely because of deep-ocean representation in POV (half of the samples in POV are from the aphotic zone). Consistent with the latter, the addition of the sole deep-ocean TOV sample, TARA_70_MESO, decreased the number of core PCs shared by all TOV samples from 220 to 65, which suggests that deep-ocean viral genetic repertoires are different from those in the upper oceans. Indeed, niche-differentiation has been observed in viromes sampled across these oceanic zones in the POV dataset (28), and similar findings were observed in the microbial metagenomic counterparts from the *Tara* Oceans Expedition (25). Thus viral communities from the deep ocean remain poorly explored and appear to hold different gene sets from those in the epipelagic oceans.

Beyond core and pan metagenomic analyses, PCs also provide a metric for viral community diversity comparisons (Fig. 3A; Supplementary File S1) from which three trends emerge in the TOV dataset. First, high-latitude viromes (82_DCM and 85_DCM) were least diverse (Shannon's H' of 8.93 and 9.22 nats), consistent with patterns in marine macroorganisms (29) and epipelagic ocean bacteria (25, 30). Second, the remaining viromes had similar diversity (Shannon's H' between 9.47 and 10.55 nats) and evenness (Pielou's J from 0.85 to 0.91) indicating low dominance of any particular PCs (31). Third, local diversity was relatively similar to global diversity (local:global ratios of H' from 0.73 to 0.87), suggesting high dispersal of viral genes (32) across the sampled ocean viral communities.

TOV Viral Populations for Assessing Global Viral Community Structure

We next estimated abundances of the 5,476 dominant viral populations in TOV, which represented up to 14.5% of aligned reads in a sample and were defined by applying empirically-derived recruitment cut-offs from naturally-occurring T4-like cyanophages (11) to high-confidence contigs from bacterial and archaeal viruses (see Methods). Assigning viral populations based on virome data remains challenging (11, 33), but here assembly of large contigs (up to 100 kb) aided our ability to accomplish not only analyses at the gene-level using PCs, but also the genome-level using viral populations. Viral populations were rarely endemic to one station (15%), and instead were commonly observed across >4 stations (47%), and up to 24 of the 26 stations (Fig. 4 and Fig. 5A). Exceptional samples include those from the Benguela upwelling region (TARA_67_SUR) and high-latitude samples from the Antarctic Circumpolar

and Falklands currents (TARA_82_DCM and TARA_85_DCM, respectively). These samples were also divergent when assessing microbial communities (TARA_82_DCM and TARA_85_DCM displayed lower microbial genetic richness; (25)) and eukaryotic communities (TARA_67_SUR had specific and unique eukaryotic communities in all size fractions; 34). While many viral populations were broadly distributed, they were much more abundant at the original location (origin inferred from longest contig assembled; see Methods) compared to alternate stations (Fig. 5B). Thus most populations were relatively widespread, but with variable sample-to-sample abundances. As was observed with PCs, diversity and evenness estimates based on viral populations were similar across all samples except for high-latitude samples (TARA_82_DCM and TARA_85_DCM) and one sample in the Red Sea (TARA_32_DCM) that displayed lower diversity (Fig. 3B; Supplementary File S1). Finally, local diversity was relatively similar to global diversity (local:global ratios of H' from 0.23 to 0.86, average 0.74, Supplementary File S1), reflecting the high dispersal of viruses as highlighted by PC analysis.

Only 39 of the 5,476 populations we identified could be affiliated to cultured viruses, reflecting the dearth of reference viral genomes in databases. These cultured viruses include those infecting the abundant and widespread hosts SAR11, SAR116, *Roseobacter*, *Prochlorococcus* and *Synechococcus* (Fig. 6). The most abundant and widespread viral populations observed in TOV lack cultured representatives (Fig. 6), which suggests that most upper ocean viruses remain to be characterized even though viruses from known dominant microbial hosts (35-39) have been cultured. Methods independent of cultivation, including viral tagging (11) and mining of microbial genomic datasets (40, 41), show promise to expand the number of available viral reference genomes (33).

Drivers of Global Viral Community Composition and Distribution

We next leveraged this global dataset to evaluate ecological drivers (including environmental variables, sample location, and microbial abundances; Supplementary File S1) of viral community structure using all three data types – morphology, populations, and PCs. These metrics revealed increasing resolution, respectively, and showed that viral community structure was influenced by region and/or environmental conditions (Table 1). We conducted the analysis of ecological drivers using all samples in this study as well as a sample subset that omitted samples with exceptional environmental conditions and divergent viral communities observed using PC and population analyses (see above; TARA_67_SUR, TARA_82_DCM, TARA_85_DCM, and TARA_70_MESO). Within the sample subset, oceanic viral communities varied significantly with Longhurst province, biome, latitude, temperature, oxygen concentration, and microbial concentrations (including total bacteria, *Synechococcus*, and *Prochlorococcus*). Viral communities were not structured by depth (surface vs DCM) except when considering PCs, likely reflecting minimal variation between samples in the epipelagic zone compared to that of globally-sourced samples, and higher resolution provided by PCs. Nutrients influenced viral community structure when considering the whole dataset, but were much less explanatory when the few high-nutrient samples were removed, except for the

influence of phosphate concentration on viral populations. Thus nutrient concentrations may influence viral community structure, but testing this hypothesis would require analysis of samples across a more continuous nutrient gradient.

Global-scale analyses of oceanic macro- (29) and micro-organisms (30) have been conducted, including a concurrent *Tara* Oceans study showing that temperature and oxygen influence microbial community structure (25). Environmental conditions have also been shown to affect global viral community morphological traits (7). Our TOV study is consistent with these earlier findings in that viral communities are influenced by temperature and oxygen concentration, but not chlorophyll concentration (Table 1). Biogeographic structuring of TOV viral communities based on the significant influence of latitude and Longhurst provinces is also consistent with the conclusion that geographic region influences community structure in Pacific Ocean viruses (42). While only PC analysis showed depth-based divergence, this likely reflects poor ($n=1$) deep sample representation in the TOV dataset as discussed above. Prior POV viral investigation and concurrent *Tara* Oceans microbial analysis, both of which have better deep-water representation, show stronger depth patterns whereby photic and aphotic zone communities diverge (25, 28, 42). Thus our results suggest biogeography of upper-ocean viral communities is structured by environmental conditions.

Since viruses require host organisms to replicate, viral community structure follows from environmental conditions shaping the host community, as observed in paired *Tara* Oceans microbial samples (25), which would then indirectly affect viral community composition. However, global distribution of viruses can also be directly influenced by environmental conditions, such as salinity, that affect their ability to infect their hosts (43). Additionally, the variable decay rates of cultivated viruses and whole viral communities (44) could also influence their distribution as viruses with lower inherent decay rates will persist for longer in the environment, and environments with more favorable conditions (such as fewer extracellular enzymes) will also contribute to increased viral persistence. Until methods to link viruses to their host cells in natural communities mature to the point of investigating this issue at larger scales (emerging possible methods reviewed by 33, 45), analyses such as ours remain the only means to assess ecological drivers of viral community structure.

To further investigate how ocean viral communities are distributed throughout the oceans, we compared population abundances between neighboring samples to assess the net direction and magnitude of population exchange (Fig. 7, see Methods). These genomic signals revealed that population exchange between dsDNA viral communities was largely directed along major oceanic current systems (46). For example, the Agulhas current and subsequent ring formation (47) connects viral communities between the Indian and Atlantic Oceans, as also observed in planktonic communities from the *Tara* Oceans expedition (48), while increased connection between the high-latitude stations (TARA_82 and TARA_85) reflects their common origin at the divergence of the Falklands and Antarctic Circumpolar currents. Further, current strength (46) was generally related to the magnitude of inter-sample population exchange, as higher and lower exchange was observed, respectively, in stronger currents such as the Agulhas current, and

within the open ocean gyres or between land-restricted basins such as the Mediterranean and Red Seas. These findings suggest that the intensity of water mass movement, in addition to environmental conditions, may explain the degree to which viral populations cluster globally (Fig. 4). Beyond such current-driven biogeographic evidence, vertical viral transport from surface to DCM samples was also observed (Fig. 4). This is consistent with POV observations wherein deep-sea viromes include a modest influx of genetic material derived from surface-ocean viruses that are presumably transported on sinking particles (28). Exceptions include areas such as the Arabian Sea upwelling region, where increased mixing and upwelling likely exceed sinking within the upper ocean.

Our TOV results enabled evaluation of a hypothesis describing the structure of viral communities in the environment. Gene-marker-based studies targeting subsets of ocean viruses previously found high local and low global diversity (49), a pattern also recently observed genome-wide in natural cyanophage populations (11). To explain this, a seed-bank viral community structure has been invoked whereby high local genetic diversity can exist by drawing variation from a common and relatively limited global gene pool (49). Our results support this hypothesis regarding viral community structure. Ecological driver analyses suggests that such local ‘seed’ communities are influenced by environmental conditions, which directly impact their microbial hosts and then indirectly restructure viral communities. These seed communities then form the ‘bank’ in neighboring samples, presumably when passively transported by ocean currents as shown here through the population-level analyses of net viral movement between samples. This systematically-sampled, global dataset suggests large- and small-scale processes play roles in structuring viral communities and offers empirical grounding for the seed-bank hypothesis with regards to viral community distribution and structure.

Conclusions

Our large-scale dataset provides a picture of global upper-ocean viral communities in which we assessed patterns using multiple parameters including morphology, populations and PCs. Our data provide advanced and complementary views on viral community structure including non-marker-gene-based diversity estimates and broad application of population-based viral ecology. We affirm the seed-bank model for viruses, hypothesized nearly a decade ago (49), which explains how high local viral diversity can be consistent with limited global diversity (11, 27). The mechanism underlying this seed-bank population structure appears to be a local production of viruses under small-scale environmental constraints and passive dispersal with oceanic currents. Improving sequencing, assembly and experimental methods are transforming the investigation of viruses in nature (33, 45), and pave the way towards assessment of viral community structure and analysis of virus-host co-occurrence networks (50) without requiring marker genes (51, 52). Such experimental and analytical progress, coupled to sampling opportunities from the *Tara* Oceans expedition, are advancing viral ecology towards the quantitative science needed to model the nano- (viruses) and micro- (microbes) scale entities driving Earth’s ecosystems.

Materials and Methods

Sample Collection

Forty-three samples were collected between November 2, 2009, and May 13, 2011, at 26 locations throughout the world's oceans (Supplementary File S1) through the *Tara* Oceans Expedition (5). These included samples from a range of depths (surface, deep chlorophyll maximum, and one mesopelagic sample) located in 7 oceans and seas, 4 different biomes and 11 Longhurst oceanographic provinces (Supplementary File S1). Longhurst provinces and biomes are defined based on Longhurst (53) and environmental features are defined based on Environment Ontology (<http://environmentontology.org/>). Sampling strategy and methodology for the *Tara* Oceans Expedition is fully described by Pesant *et al.* (54).

Environmental Parameters

Temperature, salinity, and oxygen data were collected from each station using a CTD (Sea-Bird Electronics, Bellevue, WA, USA; SBE 911plus with Searam recorder) and dissolved oxygen sensor (Sea-Bird Electronics; SBE 43). Nutrient concentrations were determined using segmented flow analysis (55) and included nitrite, phosphate, nitrite plus nitrate, and silica. Nutrient concentrations below the detection limit ($0.02 \mu\text{mol kg}^{-1}$) are reported as $0.02 \mu\text{mol kg}^{-1}$. Chlorophyll concentrations were measured using HPLC (56, 57). These environmental parameters are available in Pangaea (www.pangaea.de) using the accession numbers in Supplementary File S1.

Microbial Abundances

Flow-cytometry was used to determine the concentration of *Synechococcus*, *Prochlorococcus*, total bacteria, low-DNA bacteria, high-DNA bacteria, and the percent of bacteria with high DNA in each sample (58).

Morphological Analysis of Viral Communities

Quantitative transmission electron microscopy (qTEM) was used to evaluate the capsid diameter distributions of viral communities as previously described (7). Briefly, preserved unfiltered samples (EM-grade glutaraldehyde; Sigma-Aldrich, St. Louis, MO, USA; 2% final concentration) were flash-frozen and stored at -80°C until analysis. Viruses were deposited onto TEM grids using an air-driven ultracentrifuge (Airfuge CLS, Beckman Coulter, Brea, CA, USA), followed by positive staining of the deposited material with 2% uranyl acetate (Ted Pella, Redding, CA, USA). Samples were then examined using a transmission electron microscope (Philips CM12, FEI, Hillsboro, OR, USA) with 100 kV accelerating voltage. Micrographs of 100 viruses were collected per sample using a Macrofire Monochrome CCD camera (Optronics, Goleta, CA, USA) and analyzed using ImageJ software (US National Institutes of Health, Bethesda, MD, USA; 59) to measure the capsid diameter. A subset (21) of the 41 samples presented here had previously been analyzed in a different study (7).

Virome Construction

For each sample, 20 L of seawater were 0.22 µm-filtered and viruses were concentrated from the filtrate using iron chloride flocculation (15) followed by storage at 4°C. After resuspension in ascorbic-EDTA buffer (0.1 M EDTA, 0.2 M Mg, 0.2 M ascorbic acid, pH 6.0), viral particles were concentrated using Amicon Ultra 100 kDa centrifugal devices (Millipore), treated with DNase I (100U/mL) followed by the addition of 0.1 M EDTA and 0.1 M EGTA to halt enzyme activity, and extracted as previously described (14). Briefly, viral particle suspensions were treated with Wizard PCR Preps DNA Purification Resin (Promega, WI, USA) at a ratio of 0.5 mL sample to 1 mL resin, and eluted with TE buffer (10 mM Tris, pH 7.5, 1 mM EDTA) using Wizard Minicolumns. Extracted DNA was Covaris-sheared and size selected to 160–180 bp, followed by amplification and ligation per the standard Illumina protocol. Sequencing was done on a HiSeq 2000 system at the Genoscope facilities (Paris, France).

Quality Control of Reads and Assembly

Individual reads of 43 metagenomes were quality controlled using a combination of trimming and filtering as previously described (60). Briefly, bases were trimmed at the 5' end if the number of base calls for any base (A, T, G, C) diverged by more than two standard deviations from the average across all cycles. Conversely, bases were trimmed at the 3' end of reads if the quality score was <20. Finally, reads that were shorter than 95 bp or reads with a median quality score <20 were removed from further analyses. Assembly of reads was done using SOAPdenovo (61) where insert and k-mer size are calculated at run-time and are specific to each virome as implemented in the MOCAT pipeline (62). On average, 34.2% of the virome reads were included in the assembled contigs (min: 21.08%, max: 48.52%). Virome reads were deposited in the European Nucleotide Archive (<http://www.ebi.ac.uk/ena/>) under accession numbers reported in Supplementary File S1.

Protein Clustering

Open Reading Frames (ORFs) were predicted from all quality-controlled contigs using Prodigal (63) with default settings. Predicted ORFs were clustered based on sequence similarity as described previously (9, 10). Briefly, ORFs were initially mapped to existing clusters (POV, GOS and phage genomes), using cd-hit-2d ('-g 1 -n 4 -d 0 -T 24 -M 45000'; 60% percent identity and 80% coverage). Then the remaining, unmapped ORFs were self-clustered, using cd-hit with the same options as above. Only protein clusters (PCs) with more than two ORFs were considered *bona fide* and were used for subsequent analyses. To develop read counts per PC for statistical analyses, reads were mapped back to predicted ORFs in the contigs dataset using Mosaik with the following settings: “-a all -m all -hs 15 -minp 0.95 -mmp 0.05 -mhp 100 -act 20” (version 1.1.0021; <http://bioinformatics.bc.edu/marthlab/Mosaik>). Read counts to PCs were normalized by sequencing depth of each virome. Shannon diversity (H') was calculated from PC read counts using only PCs with more than two predicted ORFs. Observed richness is reported as

the total number of reads in each PC. Pielou's evenness (J) was calculated as the ratio of H'/H_{\max} , where $H_{\max} = \ln N$, and N = total number of observed PCs in a sample.

Analysis of Viral Populations

Considering the size of the entire dataset (3,821,756 assembled contigs), we decided to focus the analysis of viral populations using contigs most likely originating from bacterial or archaeal viruses. For this, we mined only the 22,912 contigs with more than 10 predicted genes (corresponding to an average of 6.41% of the assembled reads per sample, min: 1.29%, max: 14.52%), as the origin of contigs with only a few predicted genes can be spurious. First, we removed 6,706 contigs suspected of having originated from cellular genomes (64), whether due to free genomic DNA contamination or viral-encapsulation of cellular DNA (for example, in gene transfer agents or generalized transducing phages). These suspect cellular contigs were those containing no typical viral genes (such as virion-related genes including major capsid proteins and large subunits of the terminase) and displaying as many 'characterized genes' (such as genes with a significant similarity to a PFAM domain through Hmsearch, 65) as a typical cellular genome, whereas phage genomes are typically enriched in 'uncharacterized genes' (40). We also removed all contigs posited to originate from eukaryotic viruses. These were contigs that contained at least three predicted proteins with best BLAST hits to a eukaryotic virus, and more than half of the affiliated proteins were not associated to bacteriophages or archaeal viruses. Not surprisingly, given that eukaryotes are outnumbered by bacteria and archaea in the marine environment, this step removed only 142 contigs associated with eukaryotic viruses. From the remaining 16,124 contigs most likely to have originated from bacterial or archaeal viruses, the population study only used those longer than 10kb in size – a total of 6,322 contigs, which corresponded to an average of 4.04% of the assembled reads per sample, min: 0.98%, max: 9.97%).

These 6,322 contigs were then clustered into populations if they shared more than 80% of their genes at >95% nucleotide identity; a threshold derived from naturally-occurring T4-like cyanophages (11). This resulted in 5,476 'populations' from the 6,322 contigs, where as many as 12 contigs (average 1.15 contigs) were included per population. For each population, the longest contig was chosen as the 'seed' sequence.

The relative abundance of each population was computed by mapping all quality-controlled reads to the set of 5,476 non-redundant populations (considering only mapping quality scores greater than 1) with Bowtie 2 (66). For each sample–sequence pair, if more than 75% of the reference sequence was covered by virome reads, the relative abundance was computed as the number of base pairs recruited to the contig normalized to the total number of base pairs available in the virome and the contig length. Shannon diversity index (H') and Pielou's evenness (J) were calculated as done for PCs using the relative abundance of viral populations.

The sample containing the seed sequence (the longest contig in a population) was also considered the best estimate of that population's origin. We reasoned this was because the longest contig in a population would derive most often from the sample with the highest

coverage (a metric for population abundance) and likely corresponded to the location with the greatest viral abundance for this population. This assumption was supported by the results showing that populations were most abundant in their original samples (Fig. 4, Fig. 5B). Even though some individual cases could diverge from this rule, we expected to correctly identify most of these original locations, hence getting an accurate global signal.

The seed sequence was also used to assess taxonomic affiliation of the viral population. Cases where >50% of the genes were affiliated to a specific reference genome from RefSeq (based on a BLASTp comparison with thresholds of 50 for bit score and 10^{-5} for e-value) with an identity percentage of at least 75% (at the protein sequence level) were considered as confident affiliations to the corresponding reference virus.

Finally, estimations of net viral population movement between samples were made based on the relative abundance of populations in one sample compared to that of its neighboring samples (Fig. 4). For each neighboring sample pair, the average relative abundance of populations originating from sample A in sample B was compared with the relative abundance of populations originating from sample B in sample A. The origin of each population was defined as the sample in which the longest contig of the population was assembled. The magnitude of these differences was carried through the analysis to estimate the level of transport between each pair of samples (depicted as line width in Fig. 7) and the difference between these values was used to estimate the directionality of the transfer. For example, if sample B contains many populations from sample A, but very few populations from sample B are detected in sample A, we calculate that the net movement is from sample A to sample B. Again, while the sampling of some populations may not be strong, the net movement was calculated as the average of all shared populations between neighboring sample pairs, which corresponded to 105 different populations on average (ranging from 2 to 412).

Statistical Ordination of Samples

Viral community composition based on capsid diameter distributions (from qTEM; using 7-nm histogram bin sizes), population abundances, and normalized PC read counts (using only protein clusters with more than 20 representatives) were compared using non-metric multidimensional scaling (NMDS) performed using the ‘metaMDS’ function (default parameters) of the vegan package (67) in R version 2.15.2 (68). The influence of metadata on sample ordination was evaluated using the functions ‘envfit’ for factor variables including depth category, Longhurst province, and biome, and ‘ordisurf’ for all linear variables, in the vegan package (67, 69). Several samples had exceptional environmental conditions (TARA_67_SUR, TARA_70_MESO, TARA_82_DCM, and TARA_85_DCM), thus all statistical ordination analyses were conducted with and without these samples (referred to as the ‘sample subset’) to evaluate their influence.

References

1. C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**, 237-240 (1998).

- 463 2. P. G. Falkowski, T. Fenchel, E. F. Delong, The microbial engines that drive Earth's
464 biogeochemical cycles. *Science* **320**, 1034-1039 (2008).
- 465 3. M. Breitbart, Marine viruses: Truth or dare. *Ann. Rev. Mar. Sci.* **4**, 425-448 (2012).
- 466 4. C. A. Suttle, Marine viruses - major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**,
467 801-812 (2007).
- 468 5. E. Karsenti, *et al.*, A holistic approach to marine eco-systems biology. *PLoS Biol.* **9**, e1001177
469 (2011).
- 470 6. S. Solonenko, *et al.*, Sequencing platform and library preparation choices impact viral
471 metagenomes. *BMC Genomics* **14**, 320 (2013).
- 472 7. J. R. Brum, R. O. Schenck, M. B. Sullivan, Global morphological analysis of marine viruses
473 shows minimal regional variation and dominance of non-tailed viruses. *ISME J.* **7**, 1738-1751
474 (2013).
- 475 8. P. Hingamp, *et al.*, Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial
476 metagenomes. *ISME J.* **7**, 1678-1695 (2013).
- 477 9. S. Yooseph, *et al.*, The Sorcerer II global ocean sampling expedition: expanding the universe
478 of protein families. *PLoS Biol.* **5**, 0432-0466 (2007).
- 479 10. B. L. Hurwitz, M. B. Sullivan, The Pacific Ocean Virome (POV): A marine viral
480 metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS ONE* **8**,
481 e57355 (2013).
- 482 11. L. Deng, *et al.*, Viral tagging reveals discrete populations in *Synechococcus* viral genome
483 sequence space. *Nature* **513**, 242 (2014).
- 484 12. M. B. Duhaime, M. B. Sullivan, Ocean viruses: Rigorously evaluating the metagenomic
485 sample-to-sequence pipeline. *Virology* **434**, 181-186 (2012).
- 486 13. M. B. D. Duhaime, L. Deng, B. T. Poulos, M. B. Sullivan, Towards quantitative
487 metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous
488 assessment and optimization of the linker amplification method. *Environ. Microbiol.* **14**, 2526-
489 2537 (2012).
- 490 14. B. L. Hurwitz, L. Deng, B. T. Poulos, M. B. Sullivan, Evaluation of methods to concentrate
491 and purify ocean virus communities through comparative, replicated metagenomics. *Environ.*
492 *Microbiol.* **15**, 1428-1440 (2013).
- 493 15. S. G. John, *et al.*, A simple and efficient method for concentration of ocean viruses by
494 chemical flocculation. *Environ. Microbiol. Rep.* **3**, 195-202 (2011).
- 495 16. G. F. Steward, *et al.*, Are we missing half of the viruses in the ocean? *ISME J.* **7**, 672-679
496 (2013).
- 497 17. K. Holmfeldt, D. Odic, M. B. Sullivan, M. Middelboe, L. Riemann, Cultivated single
498 stranded DNA phages that infect marine Bacteroidetes prove difficult to detect with DNA
499 binding stains. *Appl. Environ. Microbiol.* **78**, 892-894 (2012).
- 500 18. Y. Tomaru, K. Nagasaki, Flow cytometric detection and enumeration of DNA and RNA
501 viruses infecting marine eukaryotic microalgae. *J. Oceanogr.* **63**, 215-221 (2007).
- 502 19. C. P. D. Brussaard, D. Marie, G. Bratbak, Flow cytometric detection of viruses. *J. Virol.*
503 *Methods* **85**, 175-182 (2000).
- 504 20. Y. Bettarel, T. Sime-Ngando, C. Amblard, H. Laveran, A comparison of methods for
505 counting viruses in aquatic systems. *Appl. Environ. Microbiol.* **66**, 2283-2289 (2000).
- 506 21. K. P. Hennes, C. A. Suttle, Direct counts of viruses in natural waters and laboratory cultures
507 by epifluorescence microscopy. *Limnol. Oceanogr.* **40**, 1050-1055 (1995).

- 508 22. M. G. Weinbauer, C. A. Suttle, Comparison of epifluorescence and transmission electron
509 microscopy for counting viruses in natural marine waters. *Aquat. Microb. Ecol.* **13**, 225-232
510 (1997).
- 511 23. R. T. Noble, J. A. Fuhrman, Use of SYBR Green I for rapid epifluorescence counts of marine
512 viruses and bacteria. *Aquat. Microb. Ecol.* **14**, 113-118 (1998).
- 513 24. D. Marie, C. P. D. Brussaard, R. Thyraug, G. Bratbak, D. Vaultot, Enumeration of marine
514 viruses in culture and natural samples by flow cytometry. *Appl. Environ. Microbiol.* **65**, 45-52
515 (1999).
- 516 25. S. Sunagawa, Structure and function of the global ocean microbiome. (in review).
- 517 26. D. M. Kristensen, *et al.*, Orthologous gene clusters and taxon signature genes for viruses of
518 prokaryotes. *J. Bacteriol.* **195**, 941-950 (2013).
- 519 27. C. J. Ignacio-Espinoza, S. A. Solonenko, M. B. Sullivan, The global virome: not as big as we
520 thought? *Curr. Opin. Virol.* **3**, 566-571 (2013).
- 521 28. B. L. Hurwitz, J. R. Brum, M. B. Sullivan, Depth-stratified functional and taxonomic niche
522 specialization in the 'core' and 'flexible' Pacific Ocean Virome. *ISME J.* **9**, 472-484 (2015).
- 523 29. D. P. Tittensor, *et al.*, Global patterns and predictors of marine biodiversity across taxa.
524 *Nature* **466**, 1098-1101 (2010).
- 525 30. W. J. Sul, T. A. Oliver, H. W. Ducklow, L. A. Amaral-Zettler, M. L. Sogin, Marine bacteria
526 exhibit a bipolar distribution. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 2342-2347 (2013).
- 527 31. D. I. Jarvis, *et al.*, A global perspective of the richness and evenness of traditional crop-
528 variety diversity maintained by farming communities. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 5326-
529 5331 (2008).
- 530 32. S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton
531 University Press, Princeton, NJ, 2001).
- 532 33. J. R. Brum, M. B. Sullivan, Rising to the challenge: accelerated pace of discovery transforms
533 marine virology. *Nat. Rev. Microbiol.* doi:10.1038/nrmicro3404 (2015).
- 534 34. C. de Vargas, *et al.*, Global oceans eukaryotic plankton diversity. (in review).
- 535 35. I. Kang, H.-M. Oh, D. Kang, J.-C. Cho, Genome of a SAR116 bacteriophage shows the
536 prevalence of this phage type in the oceans. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 12343-12348
537 (2013).
- 538 36. S. J. Labrie, *et al.*, Genomes of marine cyanopodoviruses reveal multiple origins of diversity.
539 *Environ. Microbiol.* **15**, 1356-1376 (2013).
- 540 37. M. B. Sullivan, *et al.*, Genomic analysis of oceanic cyanobacterial myoviruses compared
541 with T4-like myoviruses from diverse hosts and environments. *Environ. Microbiol.* **12**, 3035-
542 3056 (2010).
- 543 38. Y. Zhao, *et al.*, Abundant SAR11 viruses in the ocean. *Nature* **494**, 357-360 (2013).
- 544 39. F. Rohwer, *et al.*, The complete genomic sequence of the marine phage Roseaphage SIO1
545 shares homology with nonmarine phages. *Limnol. Oceanogr.* **45**, 408-418 (2000).
- 546 40. S. Roux, *et al.*, Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as
547 revealed by single-cell- and meta- genomics. *eLife* 10.7554/eLife.03125 (2014).
- 548 41. C. M. Mizuno, F. Rodriguez-Valera, N. E. Kimes, R. Ghai, Expanding the marine virosphere
549 using metagenomics. *PLoS Genet.* **9**, e1003987 (2013).
- 550 42. B. Hurwitz, A. Westvald, J. Brum, M. Sullivan, Modeling ecological drivers in marine viral
551 communities using comparative metagenomics and network analyses. *Proc. Natl. Acad. Sci.*
552 *U.S.A.* **111**, 10714-10719 (2014).

- 553 43. P. Kukkaro, D. H. Bamford, Virus-host interactions in environments with a wide range of
554 ionic strengths. *Environ. Microbiol. Rep.* **1**, 71-77 (2009).
- 555 44. K. E. Wommack, R. R. Colwell, Virioplankton: viruses in aquatic ecosystems. *Microbiol.*
556 *Mol. Biol. Rev.* **64**, 69-114 (2000).
- 557 45. V. Dang, M. B. Sullivan, Emerging methods to study bacteriophage infection at the single-
558 cell level. *Front. Microbiol.* (in press).
- 559 46. L. D. Talley, G. L. Pickard, W. J. Emery, J. H. Swift, *Descriptive Physical Oceanography:*
560 *An Introduction (Sixth Edition)* (Elsevier, Boston, 2011).
- 561 47. D. B. Olson, R. H. Evans, Rings of the Agulhas current. *Deep Sea Res. A* **33**, 27-42 (1986).
- 562 48. E. Villar, Dispersal and remodeling of plankton communities by Agulhas rings. (in review).
- 563 49. M. Breitbart, F. Rohwer, Here a virus, there a virus, everywhere the same virus? *Trends*
564 *Microbiol.* **13**, 278-284 (2005).
- 565 50. G. Lima-Mendez, *et al.*, Top-down determinants of ocean microbial community structure. (in
566 review).
- 567 51. D. M. Needham, *et al.*, Short-term observations of marine bacterial and viral communities:
568 patterns, connections and resilience. *ISME J.* **7**, 1274-1285 (2013).
- 569 52. C.-E. T. Chow, D. Y. Kim, R. Sachdeva, D. A. Caron, J. A. Fuhrman, Top-down controls on
570 bacterial community structure: microbial network analysis of bacteria, T4-like viruses and
571 protists. *ISME J.* **8**, 816-829 (2014).
- 572 53. A. Longhurst, *Ecological Geography of the Sea* (Elsevier, Inc., London, 2007).
- 573 54. S. Pesant, *et al.*, *Tara Oceans Data: A sampling strategy and methodology for the study of*
574 *marine plankton in their environmental context.* (in review).
- 575 55. A. Aminot, R. Kerouel, S. C. Coverly, in *Practical Guidelines for the Analysis of Seawater*,
576 O. Wurl, Eds (CRC Press, Boca Raton, 2009), vol. pp. 143-178.
- 577 56. J. Ras, H. Claustre, J. Uitz, Spatial variability of phytoplankton pigment distributions in the
578 Subtropical South Pacific Ocean: comparison between *in situ* and predicted data. *Biogeosciences*
579 **5**, 353-369 (2008).
- 580 57. L. Van Heukelem, C. S. Thomas, Computer-assisted high-performance liquid
581 chromatography method development with applications to the isolation and analysis of
582 phytoplankton pigments. *J. Chromatogr. A* **910**, 31-49 (2001).
- 583 58. J. M. Gasol, P. A. del Giorgio, Using flow cytometry for counting natural planktonic bacteria
584 and understanding the structure of planktonic bacterial communities. *Sci. Mar.* **64**, 197-224
585 (2000).
- 586 59. M. D. Abramoff, P. J. Magalhaes, S. J. Ram, Image processing with ImageJ. *Biophotonics*
587 *International* **11**, 36-42 (2004).
- 588 60. S. Schloissnig, *et al.*, Genomic variation landscape of the human gut microbiome. *Nature*
589 **493**, 45-50 (2013).
- 590 61. R. Luo, *et al.*, SOAPdenovo2: an empirically improved memory-efficient short-read de novo
591 assembler. *GigaScience* **1**, 18 (2012).
- 592 62. J. R. Kultima, *et al.*, MOCAT: A metagenomics assembly and gene prediction toolkit. *PLoS*
593 *ONE* **7**, e47656 (2012).
- 594 63. D. Hyatt, *et al.*, Prodigal: prokaryotic gene recognition and translation initiation site
595 identification. *BMC Bioinformatics* **11**, 119 (2010).
- 596 64. S. Roux, M. Krupovic, D. Debros, P. Forterre, F. Enault, Assessment of viral community
597 functional potential from viral metagenomes may be hampered by contamination with cellular
598 sequences. *Open Biol.* **3**, 130160 (2013).

65. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29-W37 (2011).
66. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359 (2012).
67. J. Oksanen, *et al.*, vegan: Community Ecology Package, R package version 2.1-27/r2451 (2013).
68. R Core Team, R: A language and environment for statistical computing. v. 2.15.2 (2012).
69. S. N. Wood, Fast stable restricted maximum likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. A* **73**, 3-36 (2011).

Acknowledgements. We thank Jesse Czekanski-Moir for advice on statistics and Laurent Coppola for assistance with validating nutrient data. We thank the commitment of the following people and sponsors: CNRS (in particular Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, Fund for Scientific Research – Flanders, Rega Institute, KU Leuven, The French Ministry of Research, the French Government 'Investissements d'Avenir' programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001-02), ANR (projects POSEIDON/ANR-09-BLAN-0348, PHYTBACK/ANR-2010-1709-01, PROMETHEUS/ANR-09-PCS-GENM-217, TARA-GIRUS/ANR-09-PCS-GENM-218), European Union FP7 (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376), ERC Advanced Grant Award to CB (Diatomite: 294823), Gordon and Betty Moore Foundation grant (#3790) to MBS, Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS to SGA, TANIT (CONES 2010-0036) from the Agència de Gestió d'Ajuts Universitaris i Reserca to SGA, JSPS KAKENHI Grant Number 26430184 to HO, and FWO, BIO5, Biosphere 2 to MBS. We also thank the support and commitment of Agnès b. and Etienne Bourgois, the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, the *Tara* schooner and its captains and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who graciously granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (<http://oceans.taraexpeditions.org>). We also acknowledge excellent assistance from the European Bioinformatics Institute (EBI), in particular Guy Cochrane and Petra ten Hoopen, as well as the EMBL Advanced Light Microscopy Facility (ALMF), in particular Rainer Pepperkok. The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled in. Data described herein is available at EBI (project identifiers PRJEB402 and PRJEB7988) and Pangaea (see Supplementary Table S1), and the data release policy regarding future public release of *Tara* Oceans data is described in Pesant *et al.* (54). We also acknowledge the EMBL Advanced Light Microscopy Facility (ALMF), and in particular Rainer Pepperkok. This research is funded in part by the Gordon and Betty Moore Foundation through grants GBMF2631 and GBMF3790 to MBS. We also acknowledge support from UA high-performance computing; the foundation for France-American Cultural Exchange, Partner University Fund program, awarded to Ecole

646 Normale Supérieure and the University of Arizona; and a grant to the University of Arizona
647 Ecosystem Genomics Institute through the UA Technology and Research Initiative Fund and the
648 Water, Environmental and Energy Solutions Initiative. All authors approved the final
649 manuscript. This article is contribution number XXX of the *Tara* Oceans Expedition.
650 Supplement contains additional data.

651 ***Tara* Oceans Coordinators**

652 Silvia G. Acinas¹, Peer Bork^{2,3}, Emmanuel Boss⁴, Chris Bowler⁵, Colombar de Vargas^{6,7},
653 Michael Follows⁸, Gabriel Gorsky^{9,31}, Nigel Grimsley^{10,11}, Pascal Hingamp¹², Daniele
654 Iudicone¹³, Olivier Jaillon^{14,15,16}, Stefanie Kandels-Lewis^{2,17}, Lee Karp-Boss¹⁸, Eric Karsenti^{5,17},
655 Uros Krzic¹⁹, Fabrice Not^{6,7}, Hiroyuki Ogata²⁰, Stephane Pesant^{21,22}, Jeroen Raes^{23,24,25},
656 Emmanuel G. Reynaud²⁶, Christian Sardet^{27,28}, Mike Sieracki^{29,†}, Sabrina Speich^{30,‡}, Lars
657 Stemmann^{9,31}, Matthew B. Sullivan³², Shinichi Sunagawa², Didier Velayoudon³³, Jean
658 Weissenbach^{14,15,16}, Patrick Wincker^{14,15,16}

659 ¹ Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)-CSIC,
660 Pg. Marítim de la Barceloneta 37-49, Barcelona, E08003, Spain

661 ² Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr.
662 1, 69117 Heidelberg, Germany

663 ³ Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany

664 ⁴ School of Marine Sciences, University of Maine, Orono, Maine, USA

665 ⁵ Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and
666 CNRS UMR 8197, Paris, F-75005 France

667 ⁶ CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff,
668 France

669 ⁷ Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Place
670 Georges Teissier, 29680 Roscoff, France

671 ⁸ Dept of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology,
672 Cambridge, Massachusetts, USA

673 ⁹ CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France

674 ¹⁰ CNRS UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France

675 ¹¹ Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer,
676 France

677 ¹² Aix Marseille Université CNRS IGS UMR 7256 13288 Marseille, France

678 ¹³ Stazione Zoologica Anton Dohrn, Villa Comunale, 80121, Naples, Italy

- 679 ¹⁴ CEA - Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France
- 680 ¹⁵ CNRS, UMR 8030, CP5706, Evry, France
- 681 ¹⁶ Université d'Evry, UMR 8030, CP5706, Evry, France
- 682 ¹⁷ Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1 69117
683 Heidelberg, Germany
- 684 ¹⁸ School of Marine Sciences, University of Maine, Orono, USA
- 685 ¹⁹ Cell Biology and Biophysics, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117
686 Heidelberg, Germany
- 687 ²⁰ Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-001, Japan
- 688 ²¹ PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen,
689 Bremen, Germany
- 690 ²² MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen,
691 Germany
- 692 ²³ Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49,
693 3000 Leuven, Belgium
- 694 ²⁴ Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium
- 695 ²⁵ Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050
696 Brussels, Belgium
- 697 ²⁶ Earth Institute, University College Dublin, Dublin, Ireland
- 698 ²⁷ CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer,
699 France
- 700 ²⁸ Sorbonne Universités, UPMC Univ Paris 06, UMR 7009 Biodev, F-06230 Observatoire
701 Océanologique, Villefranche-sur-mer, France
- 702 ²⁹ Bigelow Laboratory for Ocean Science, East Boothbay, Maine, USA
- 703 ³⁰ Laboratoire de Physique des Océan UBO-IUEM Palce Copernic 29820 Polouzané, France
- 704 ³¹ Sorbonne Universités, UPMC Univ Paris 06, UMR 7093, LOV, Observatoire Océanologique,
705 F-06230, Villefranche-sur-mer, France
- 706 ³² Department of Ecology and Evolutionary Biology, Depts Molecular and Cellular Biology and
707 Soil, Water and Environmental Science, University of Arizona, Tucson, Arizona, 85721, USA
- 708 ³³ DVIP Consulting, 92310, Sèvres, France

[†] Current address: National Science Foundation, Arlington, Virginia, USA

[‡] Current address: Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD),
Ecole Normale Supérieure, 24 rue Lhomond 75231 Paris, Cedex 05, France

Supplementary Information

Supplementary File S1. Description of samples and relevant virome data. Metadata is presented for each *Tara* Oceans sample in this study including the PANGAEA accession numbers, sample location and environmental conditions, and the abundances of selected microorganisms. Detailed information is also presented for the viromes in this study including ENA accession numbers, the total number of reads and PCs for each virome, and diversity and evenness data for each virome based on PCs and viral populations.

Figure Legends

Fig. 1. Distribution of viral capsid diameters in each sample (n = 100 viruses per sample). Data are not available for samples TARA_18_DCM and TARA_70_MESO. Boxplots are constructed with the upper and lower lines corresponding to the 25th and 75th percentiles, while outliers are displayed as points. Longhurst provinces are indicated below samples (MEDI, Mediterranean Sea; REDS, Red Sea; ARAB, NW Arabian Upwelling; MONS, Indian Monsoon Gyres; ISSG, Indian S. Subtropical Gyre; EAFR, E. Africa Coastal; BENG, Benguela Current Coastal; SATL, S. Atlantic Gyre; FKLD, SW Atlantic Shelves; APLR, Austral Polar; PNEC, N. Pacific Equatorial Countercurrent).

Fig. 2. Protein cluster (PC) richness in core and pan viromes from the TOV and POV datasets. A) Accumulation curves of core and pan PCs in the TOV dataset. Vertical axis shows the number of shared (core virome) and total (pan virome) PCs when *n* viromes are compared (*n* = 1 to 43; from 3 to 41 only 1000 combinations are shown). Lines: i) total number of PCs (1,075,763 PCs), ii) core surface virome (710 PCs), iii) core DCM virome (424 PCs), iv) core surface and DCM virome (220 PCs), v) all samples (including the deep-ocean sample TARA_70_MESO; 65 PCs). B) Core and pan PCs in all TOV and photic-zone POV samples combined. Vertical axis shows the number of shared (core virome) and total (pan-virome) PCs when *n* viromes are compared (*n* = 1 to 57; from 3 to 57 only 1,000 combinations are shown). Overall, 1,323,921 PCs were identified in all viromes combined.

Fig. 3. Alpha diversity measurements in TOV dataset. A) Shannon's richness *H'* and Pielou's evenness *J* calculated from protein clusters counts for each sample and a pool of all samples, normalized to 5 million reads. B) Shannon's richness *H'* and Pielou's evenness *J* calculated from relative abundances of viral populations for each sample and a pool of all samples, with subsamples of 100,000 reads. Outliers corresponding to values outside of the average value plus or minus two standard deviations are colored in green and red, respectively. Values calculated

from the pool of all samples are colored in blue. Longhurst provinces are indicated below samples using the same abbreviations as in Fig. 1.

Fig. 4. Relative abundance of viral populations in TOV by sample. This heatmap displays the relative abundance of each population (sorted according to its original sample; y-axis) in each sample (x-axis). Relative abundance of one population in a sample is based on recruitment of reads to the population reference contig, and only considered if more than 75% of the reference contig is covered. Longhurst provinces are indicated below samples (using the same abbreviations as in Fig. 1) and outlined in black on the heatmap.

Fig. 5. Relative abundance of viral populations in TOV by station. A) Evaluation of viral population distribution showing the number of stations (y-axis) in which each population (sorted by their original station, x-axis) is distributed. Populations are grouped by station, merging surface and DCM samples from the same station. B) Relative abundance of populations at the original stations where the contigs were assembled compared to their abundance at other stations. Boxplots are constructed as in Fig. 1.

Fig. 6. Taxonomic affiliation of TOV viral populations sorted by distribution and average abundance. A population was considered as similar to a known virus when less than half of its reference contig genes were uncharacterized, and all characterized genes had taxonomic affiliations to the same reference genome. As in Fig. 4, the relative abundance (y-axis) is computed for each sample as the number of bp mapped to a contig per kb of contig per Mb of metagenome sequenced. Here, the relative abundance of a population is defined as the average abundance of its reference contig across all samples.

Fig. 7. Net movement of viral populations throughout the oceans. Calculations are based on reciprocal comparison of viral population abundances between neighboring samples (see Fig. 3 and Methods). For each sample pair, the average relative population abundances in one sample originating from a neighboring sample were calculated and compared (for example, relative abundance of populations from sample A found in sample B are compared with relative abundance of populations from sample B found in sample A). The sign of the relative abundance difference between neighboring samples was used to estimate the movement direction (arrowhead), and the absolute value of the difference was interpreted as reflecting the movement magnitude (line width). Stations are labeled with station number. ‘Down’ and ‘up’ refer to net vertical movement of viral populations between the surface and DCM samples at the same station.

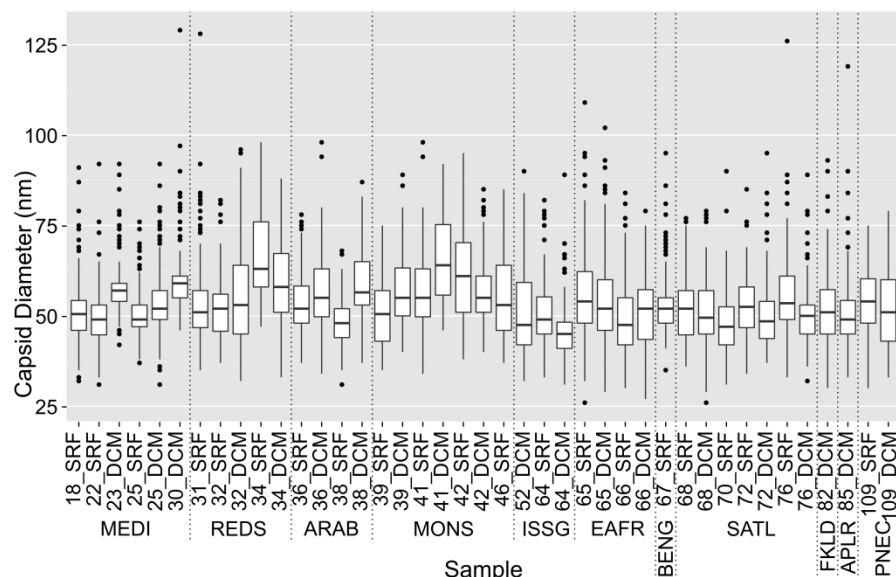


Fig. 1. Distribution of viral capsid diameters in each sample (n = 100 viruses per sample). Data are not available for samples TARA_18_DCM and TARA_70_MESO. Boxplots are constructed with the upper and lower lines corresponding to the 25th and 75th percentiles, while outliers are displayed as points. Longhurst provinces are indicated below samples (MEDI, Mediterranean Sea; REDS, Red Sea; ARAB, NW Arabian Upwelling; MONS, Indian Monsoon Gyres; ISSG, Indian S. Subtropical Gyre; EAFR, E. Africa Coastal; BENG, Benguela Current Coastal; SATL, S. Atlantic Gyre; FKLD, SW Atlantic Shelves; APLR, Austral Polar; PNEC, N. Pacific Equatorial Countercurrent).

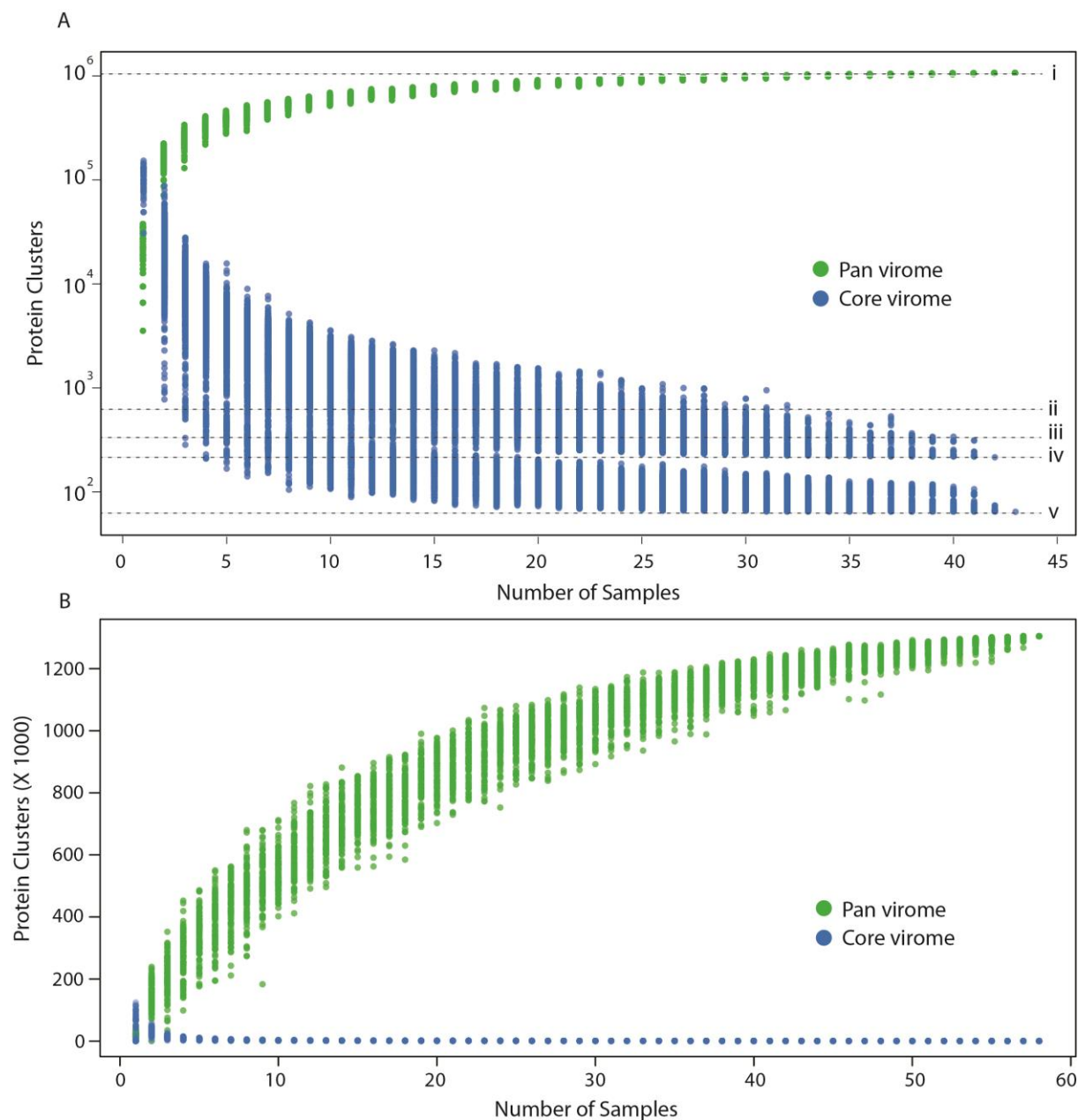


Fig. 2. Protein cluster (PC) richness in core and pan viromes from the TOV and POV datasets. A) Accumulation curves of core and pan PCs in the TOV dataset. Vertical axis shows the number of shared (core virome) and total (pan virome) PCs when n viromes are compared ($n = 1$ to 43; from 3 to 41 only 1000 combinations are shown). Lines: i) total number of PCs (1,075,763 PCs), ii) core surface virome (710 PCs), iii) core DCM virome (424 PCs), iv) core surface and DCM virome (220), v) all samples (including the deep-ocean sample TARA_70_MESO; 65 PCs). B) Core and pan PCs in all TOV and photic-zone POV samples combined. Vertical axis shows the number of shared (core virome) and total (pan-virome) PCs when n viromes are compared ($n = 1$ to 57; from 3 to 57 only 1,000 combinations are shown). Overall, 1,323,921 PCs were identified in all viromes combined.

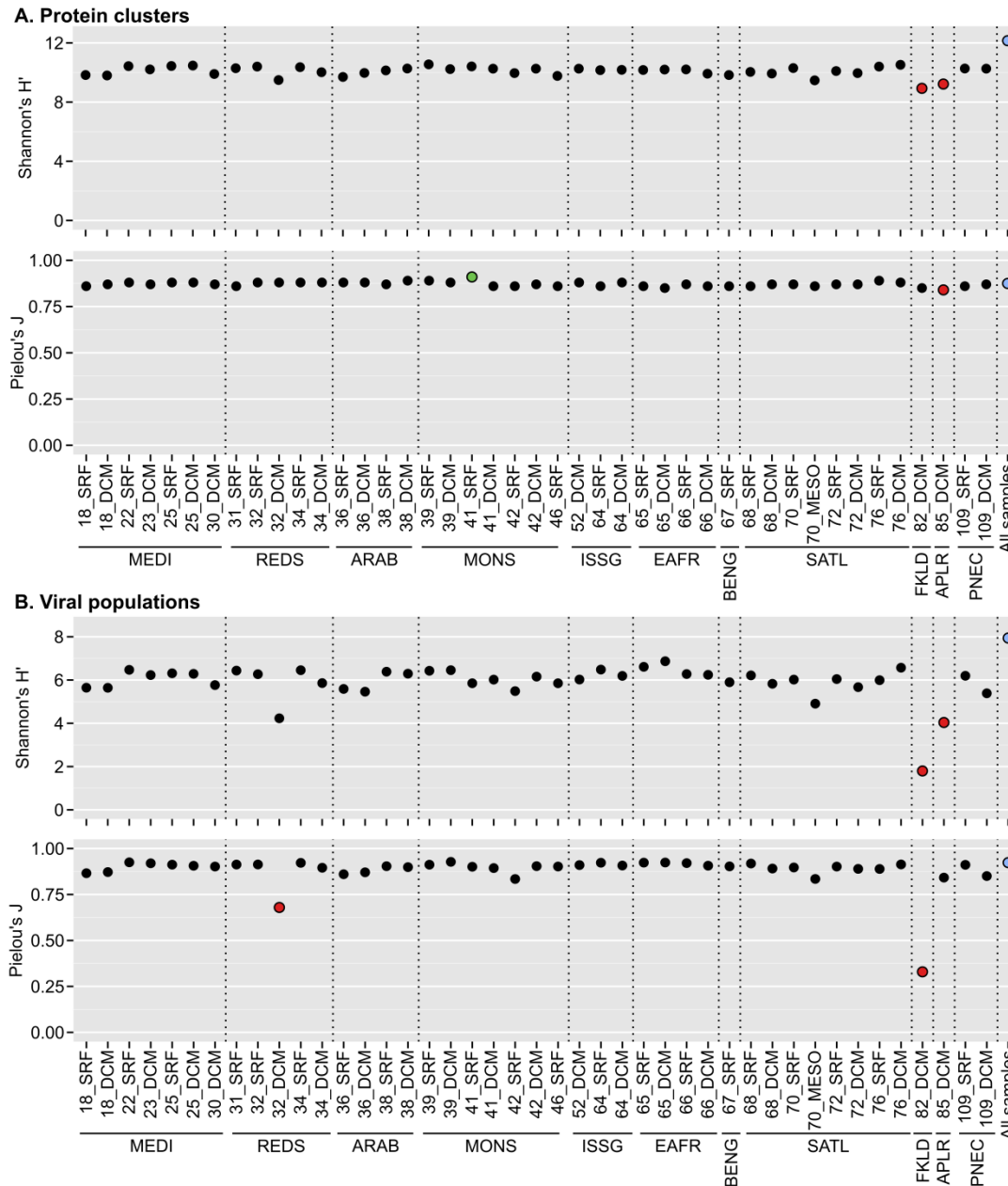


Fig. 3. Alpha diversity measurements in TOV dataset. A) Shannon's richness H' and Pielou's evenness J calculated from protein clusters counts for each sample and a pool of all samples, normalized to 5 million reads. B) Shannon's richness H' and Pielou's evenness J calculated from relative abundances of viral populations for each sample and a pool of all samples, with subsamples of 100,000 reads. Outliers corresponding to values outside of the average value plus or minus two standard deviations are colored in green and red, respectively. Values calculated from the pool of all samples are colored in blue. Longhurst provinces are indicated below samples using the same abbreviations as in Fig. 1.

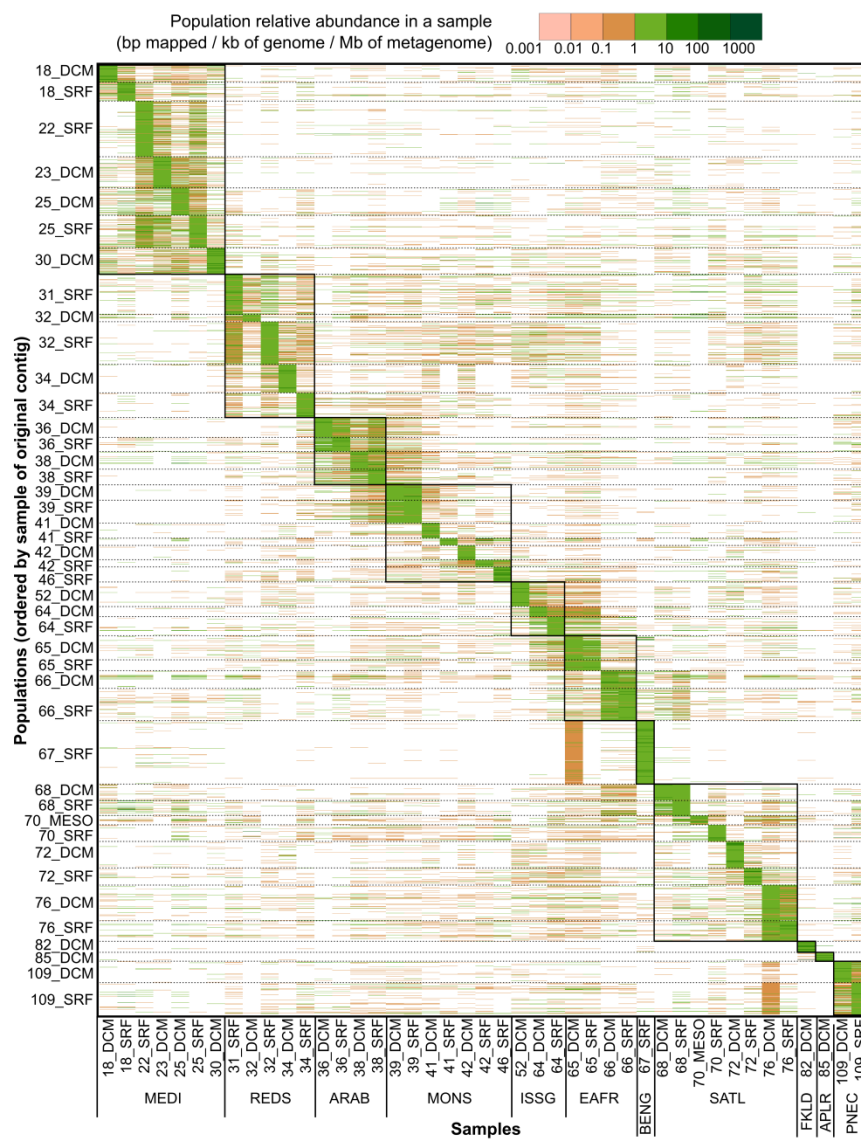


Fig. 4. Relative abundance of viral populations in TOV by sample. This heatmap displays the relative abundance of each population (sorted according to its original sample; y-axis) in each sample (x-axis). Relative abundance of one population in a sample is based on recruitment of reads to the population reference contig, and only considered if more than 75% of the reference contig is covered. Longhurst provinces are indicated below samples (using the same abbreviations as in Fig. 1) and outlined in black on the heatmap.

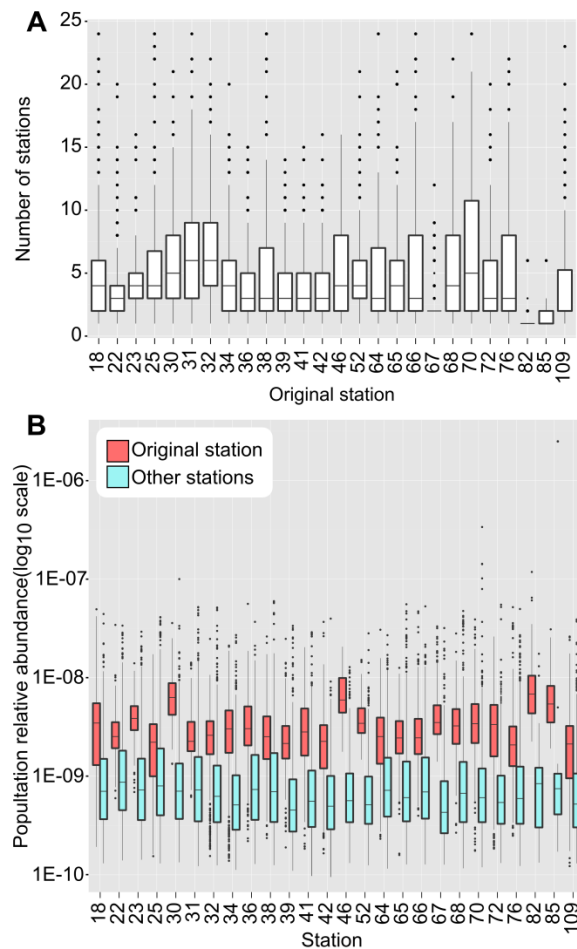


Fig. 5. Relative abundance of viral populations in TOV by station. A) Evaluation of viral population distribution showing the number of stations (y-axis) in which each population (sorted by their original station, x-axis) is distributed. Populations are grouped by station, merging surface and DCM samples from the same station. B) Relative abundance of populations at the original stations where the contigs were assembled compared to their abundance at other stations. Boxplots are constructed as in Fig. 1.

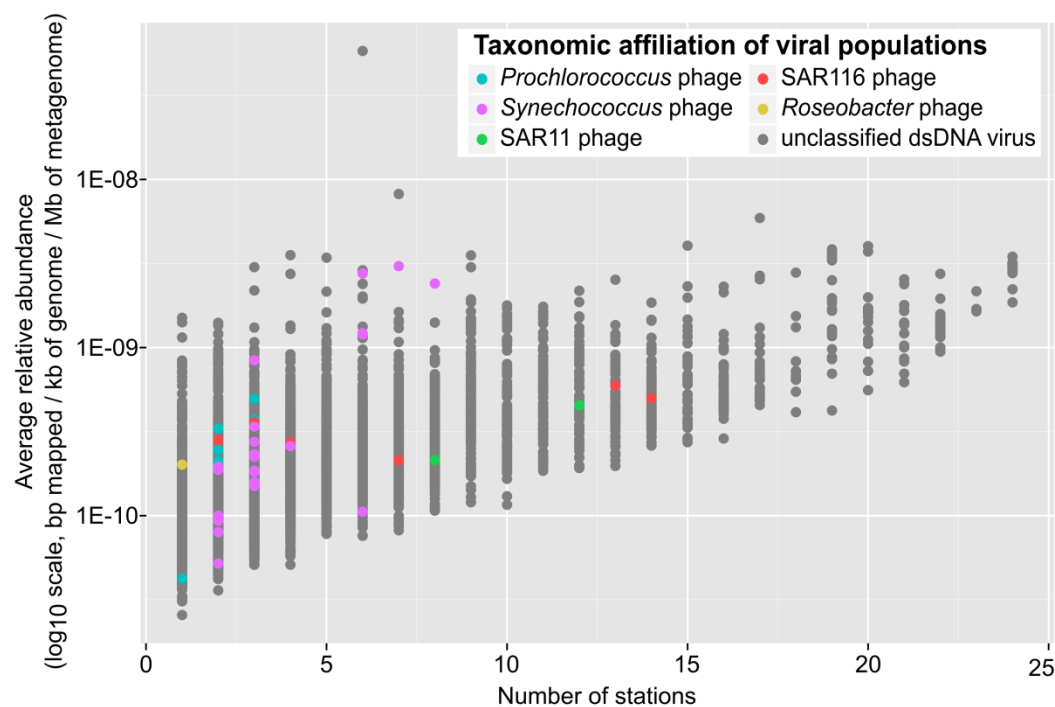


Fig. 6. Taxonomic affiliation of TOV viral populations sorted by distribution and average abundance. A population was considered as similar to a known virus when less than half of its reference contig genes were uncharacterized, and all characterized genes had taxonomic affiliations to the same reference genome. As in Fig. 4, the relative abundance (y-axis) is computed for each sample as the number of bp mapped to a contig per kb of contig per Mb of metagenome sequenced. Here, the relative abundance of a population is defined as the average abundance of its reference contig across all samples.

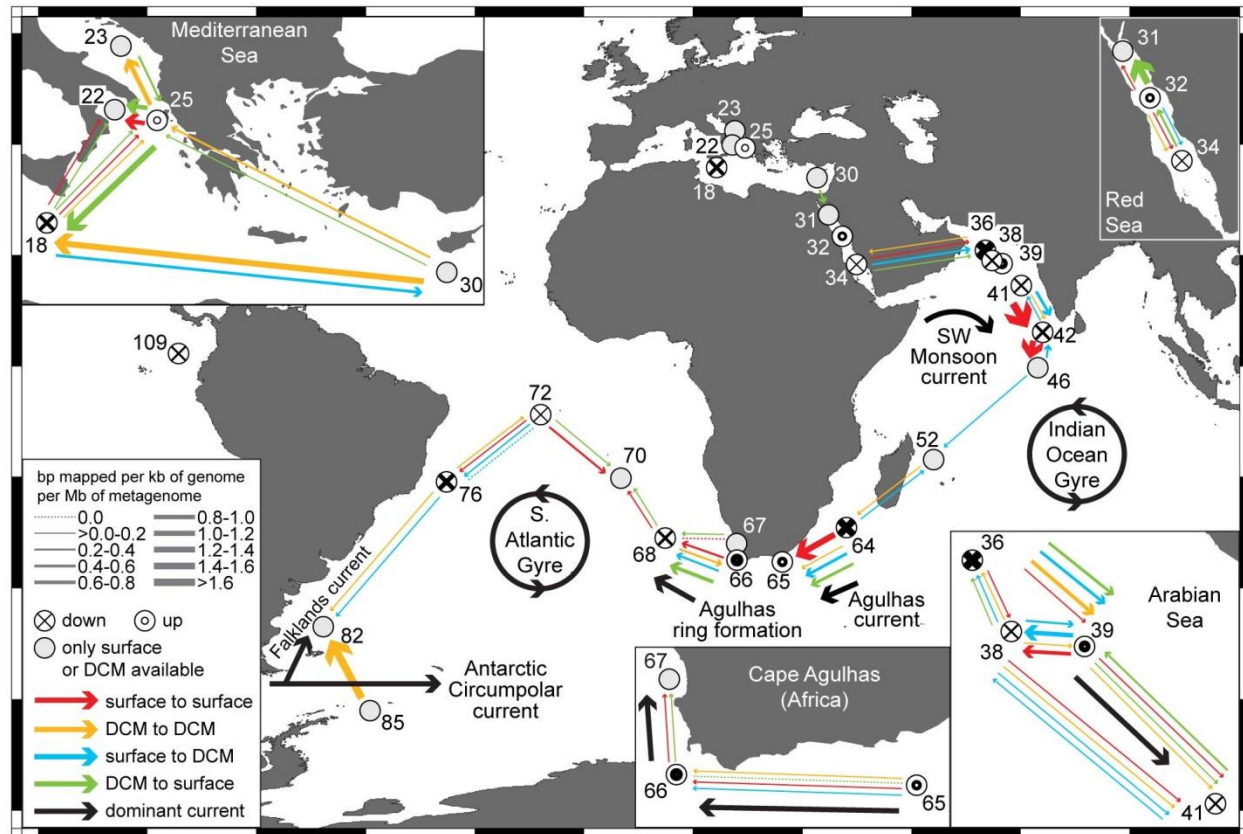


Fig. 7. Net movement of viral populations throughout the oceans. Calculations are based on reciprocal comparison of viral population abundances between neighboring samples (see Fig. 3 and Methods). For each sample pair, the average relative population abundances in one sample originating from a neighboring sample were calculated and compared (for example, relative abundance of populations from sample A found in sample B are compared with relative abundance of populations from sample B found in sample A). The sign of the relative abundance difference between neighboring samples was used to estimate the movement direction (arrowhead), and the absolute value of the difference was interpreted as reflecting the movement magnitude (line width). Stations are labeled with station number. ‘Down’ and ‘up’ refer to net vertical movement of viral populations between the surface and DCM samples at the same station.

Table 1. Relationships between viral community structure (based on viral morphology, populations, and PCs) and metadata using NMDS analysis of all samples and the sample subset (all samples except for TARA_67_SRF, TARA_70_MESO, TARA_82_DCM, and TARA_85_DCM due to exceptional environmental conditions at these locations). Significant relationships are italicized and in bold.

		Viral Morphology (qTEM)	Populations (contigs)	Protein Clusters (PCs)
Depth Category	all samples	p = 0.354 (n = 41)	p = 0.362 (n = 43)	<i>p = 0.033 (n = 43)</i>
	sample subset	p = 0.228 (n = 38)	p = 0.105 (n = 39)	<i>p = 0.011 (n = 39)</i>
Province	all samples	p = 0.098 (n = 41)	<i>p < 0.001 (n = 43)</i>	<i>p = 0.014 (n = 43)</i>
	sample subset	<i>p = 0.029 (n = 38)</i>	<i>p < 0.001 (n = 39)</i>	<i>p = 0.008 (n = 39)</i>
Biome	all samples	p = 0.099 (n = 41)	<i>p < 0.001 (n = 43)</i>	p = 0.097 (n = 43)
	sample subset	p = 0.120 (n = 38)	<i>p < 0.001 (n = 39)</i>	p = 0.543 (n = 39)
Latitude	all samples	<i>p = 0.003 (n = 41)</i>	<i>p < 0.001 (n = 43)</i>	<i>p = 0.002 (n = 43)</i>
	sample subset	<i>p = 0.014 (n = 38)</i>	<i>p < 0.001 (n = 39)</i>	<i>p = 0.010 (n = 39)</i>
Temperature	all samples	<i>p = 0.001 (n = 41)</i>	<i>p < 0.001 (n = 43)</i>	<i>p < 0.001 (n = 43)</i>
	sample subset	<i>p = 0.001 (n = 38)</i>	<i>p < 0.001 (n = 39)</i>	<i>p = 0.015 (n = 39)</i>
Salinity	all samples	p = 0.118 (n = 39)	<i>p = 0.035 (n = 41)</i>	<i>p = 0.029 (n = 41)</i>
	sample subset	p = 0.138 (n = 36)	p = 0.075 (n = 37)	<i>p = 0.001 (n = 37)</i>
Oxygen	all samples	<i>p = 0.001 (n = 41)</i>	<i>p < 0.001 (n = 43)</i>	<i>p < 0.001 (n = 43)</i>
	sample subset	<i>p = 0.005 (n = 38)</i>	<i>p < 0.001 (n = 39)</i>	<i>p < 0.001 (n = 39)</i>
Chlorophyll	all samples	p = 0.711 (n = 41)	<i>p < 0.001 (n = 43)</i>	<i>p = 0.001 (n = 39)</i>
	sample subset	p = 0.738 (n = 38)	p = 0.412 (n = 39)	p = 0.059 (n = 39)
Nitrite	all samples	p = 0.951 (n = 39)	p = 0.648 (n = 41)	p = 0.828 (n = 41)
	sample subset	p = 0.851 (n = 36)	p = 0.509 (n = 37)	p = 0.999 (n = 37)
Phosphate	all samples	p = 0.275 (n = 39)	<i>p < 0.001 (n = 41)</i>	<i>p < 0.001 (n = 41)</i>
	sample subset	p = 0.411 (n = 36)	<i>p < 0.001 (n = 37)</i>	p = 0.583 (n = 37)
Nitrite+Nitrate	all samples	<i>p = 0.046 (n = 39)</i>	<i>p < 0.001 (n = 41)</i>	<i>p < 0.001 (n = 41)</i>
	sample subset	p = 0.290 (n = 36)	p = 0.052 (n = 37)	p = 0.643 (n = 37)
Silica	all samples	<i>p = 0.008 (n = 39)</i>	<i>p = 0.002 (n = 41)</i>	<i>p = 0.008 (n = 41)</i>
	sample subset	p = 0.255 (n = 36)	p = 0.285 (n = 37)	p = 0.191 (n = 37)
Bacteria	all samples	p = 0.579 (n = 39)	<i>p < 0.001 (n = 40)</i>	p = 0.119 (n = 40)
	sample subset	p = 0.329 (n = 36)	<i>p = 0.003 (n = 36)</i>	<i>p = 0.007 (n = 36)</i>
Low DNA bacteria	all samples	p = 0.227 (n = 39)	p = 0.090 (n = 40)	p = 0.123 (n = 40)
	sample subset	p = 0.468 (n = 36)	<i>p = 0.018 (n = 36)</i>	<i>p = 0.005 (n = 36)</i>
High DNA bacteria	all samples	p = 0.967 (p = 39)	<i>p < 0.001 (n = 40)</i>	p = 0.273 (n = 40)
	sample subset	p = 0.174 (n = 36)	<i>p = 0.027 (n = 36)</i>	<i>p = 0.024 (n = 36)</i>
% high DNA bacteria	all samples	<i>p = 0.007 (n = 39)</i>	p = 0.078 (n = 40)	<i>p = 0.009 (n = 40)</i>
	sample subset	<i>p = 0.017 (n = 36)</i>	p = 0.059 (n = 36)	<i>p < 0.001 (n = 36)</i>
<i>Synechococcus</i>	all samples	p = 0.143 (n = 39)	p = 0.094 (n = 40)	<i>p = 0.041 (n = 40)</i>
	sample subset	p = 0.142 (n = 36)	<i>p = 0.023 (n = 36)</i>	<i>p = 0.013 (n = 36)</i>
<i>Prochlorococcus</i>	all samples	p = 0.118 (n = 39)	p = 0.076 (n = 40)	p = 0.123 (n = 40)
	sample subset	p = 0.249 (n = 37)	p = 0.161 (n = 37)	p = 0.140 (n = 37)

[illegible]

* Values below detection limit are reported as the detection limit of 0.01 $\mu\text{mol kg}^{-1}$.

[illegible]